Polish company Bankruptcy data analysis

The dataset

The dataset is about bankruptcy prediction of Polish companies. The data was collected from Emerging Markets Information Service (EMIS, [Web Link]), which is a database containing information on emerging markets around the world. The bankrupt companies were analysed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013.

The dataset

- Contains 5 file for five different year
- Contains 64 attributes and the bankruptcy status
- Basing on the collected data five classification cases were distinguished, that depends on the forecasting period, all files contains financial informations but also:
  - - 1stYear bankruptcy status after 5 years.
  - - 2ndYear bankruptcy status after 4 years.
  - - 3rdYear bankruptcy status after 3 years.
  - - 4thYear bankruptcy status after 2 years.
  - - 5thYear bankruptcy status after 1 years.

The dataset

We need to determine which attribute is a real indicator of the health of a company. It is a classification problem.

We can rely on classifier algorithm to do this job with binary logistic objective.

To measure the performance of this model I will use AUC indicator due to the imbalanced nature of the data.

Usage

This dataset can be very useful to determine the health of a company.

By using each dataset we can predict if the company go bankrupt in 1 year, 2 year…

My test

- I did my test on the 1stYear dataset
- First I try all Sklearn implemented algorithms plus XGboost and catboost

Top three algorithm:

C-Support Vector Classification	0.991242
Quadratic Discriminant Analysis	0.991242
Xgboost	0.991242

They had an equal score, I choose XGboost because of is tunable features

XGboost optimisation

- I tune xgboost parameters:
  - max_depth
  - min_child_weight
  - Subsample
  - eta

API

I add an API to predict if a company will go bankrupt.

The endpoint is:

http://127.0.0.1:/is_bankrupt